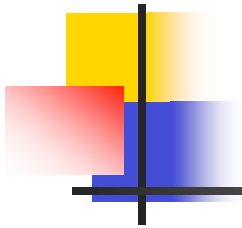
GDS : Grid Data Service

Un service de gestion de données pour les grilles de calcul

Gabriel Antoniu[1], Eddy Caron[2],
Frédéric Desprez[2], Pierre Sens[3]

[1]IRISA, Rennes
[2]LIP, ENS Lyon
[3]LIP6, Université Paris 6

Journées PaRISTIC, Bordeaux, 21-23 novembre 2005



Le projet GDS

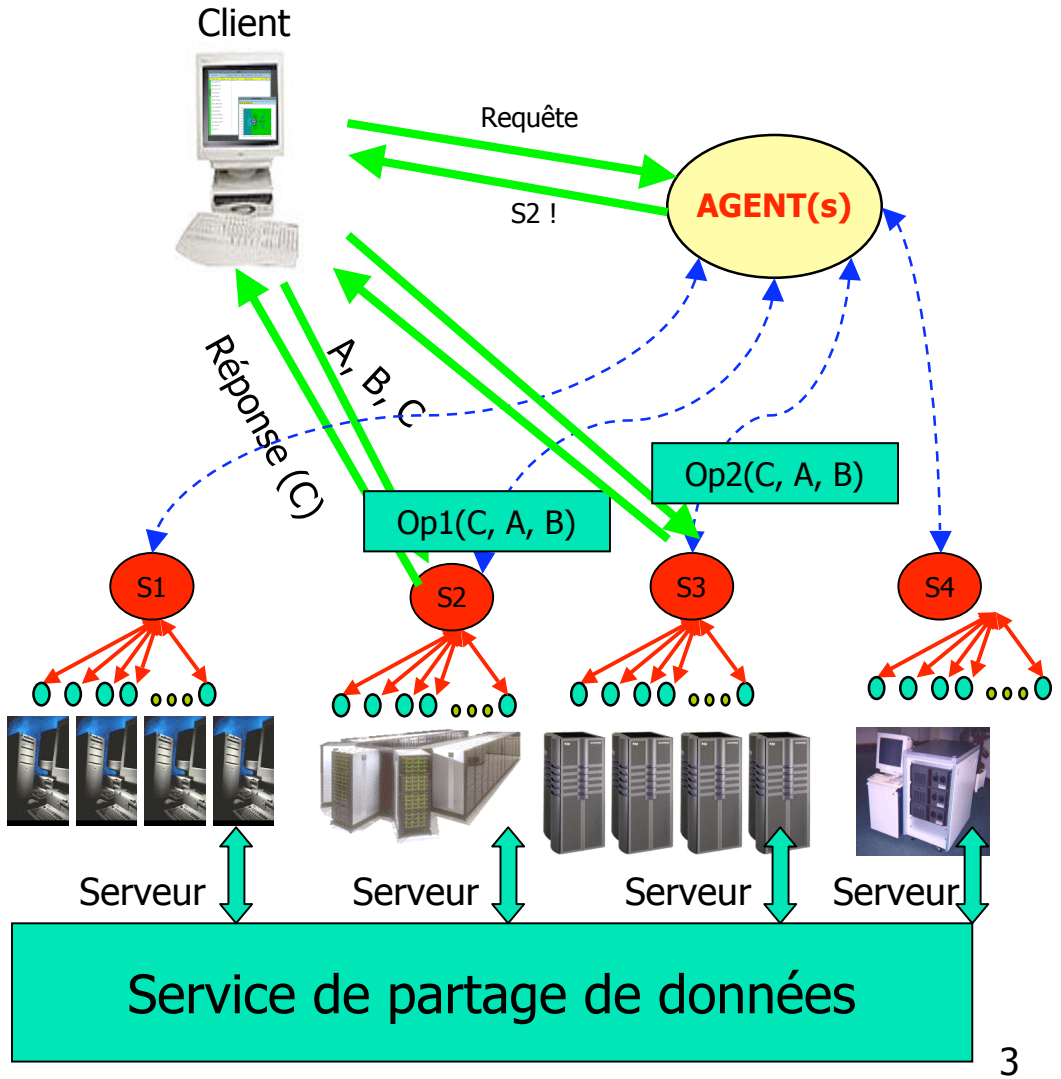
- Projet de l'ACI Masses de Données (2003)
- Objectif :

Spécifier, réaliser et valider expérimentalement un service de partage de données pour la grille, adapté au calcul scientifique

- Partenaires
  - Projet GRAAL (LIP, Lyon)
  - Projet PARIS (IRISA, Rennes)
  - Projet REGAL (LIP6, Paris)

GDS : objectifs

- Partager des données entre plusieurs composants applicatifs distribués sur une grille
  - Transparence de la localisation des données
  - Stockage persistant
  - Tolérance aux fautes
  - Gestion de la cohérence des copies en environnement dynamique
  - Extensibilité à l'échelle d'une grille : des milliers de nœuds

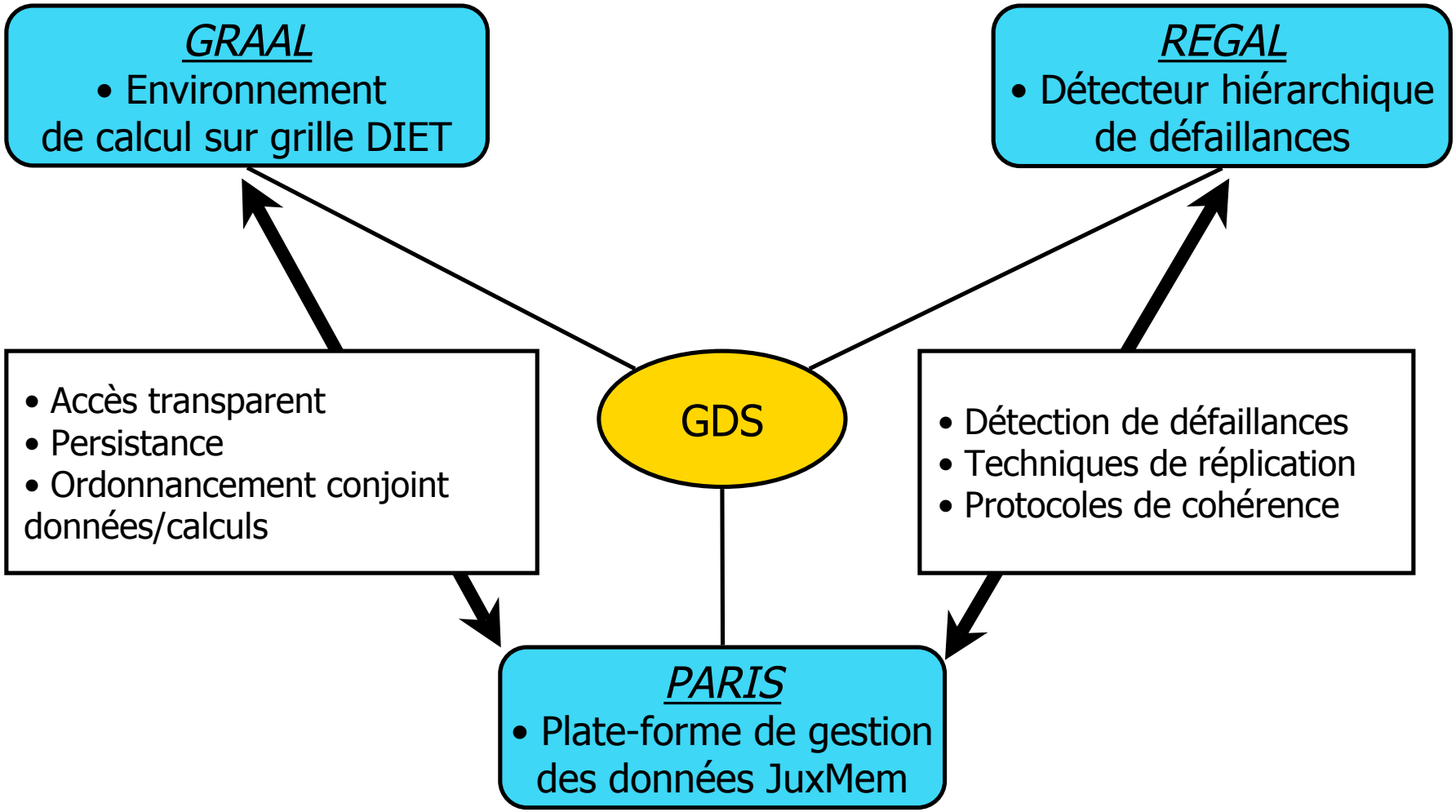


Idée : une approche hybride

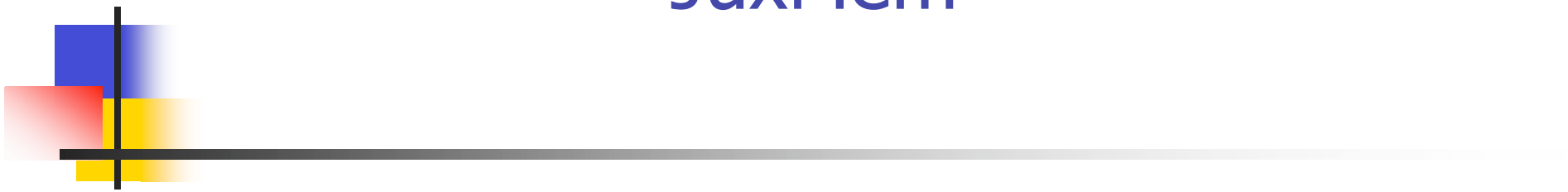
- Systèmes MVP (DSM) : cohérence et accès transparent
- Systèmes P2P : extensibilité et volatilité

	MVP	**Service de partage pour la grille**	**P2P**
Plate-forme physique	Grappe	Fédération de grappes	Internet
Echelle	10^1-10^2	10^3 - 10^4	10^5-10^6
Volatilité	Nulle	Moyenne	Haute
Contrôle	Fort	Moyen	Faible
Applications	Calcul scientifique	Calcul scientifique et stockage	Partage et stockage de fichiers
Type de données	Modifiables	Modifiables	Non modifiables
Tolérance aux fautes	Rare	?	Critique
Cohérence	Critique	?	Rare

GDS : interaction des partenaires

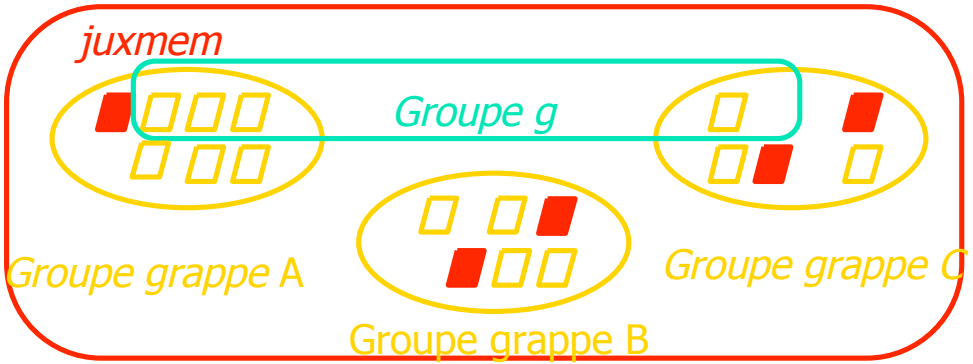


JuxMem



JuxMem

- Une architecture pair-à-pair de service de partage de données en mémoire
- Mécanisme de persistance
- Localisation transparente des données
- Mécanisme de réplication
- Mécanismes de cohérence



Architecture logique

Project JXTA

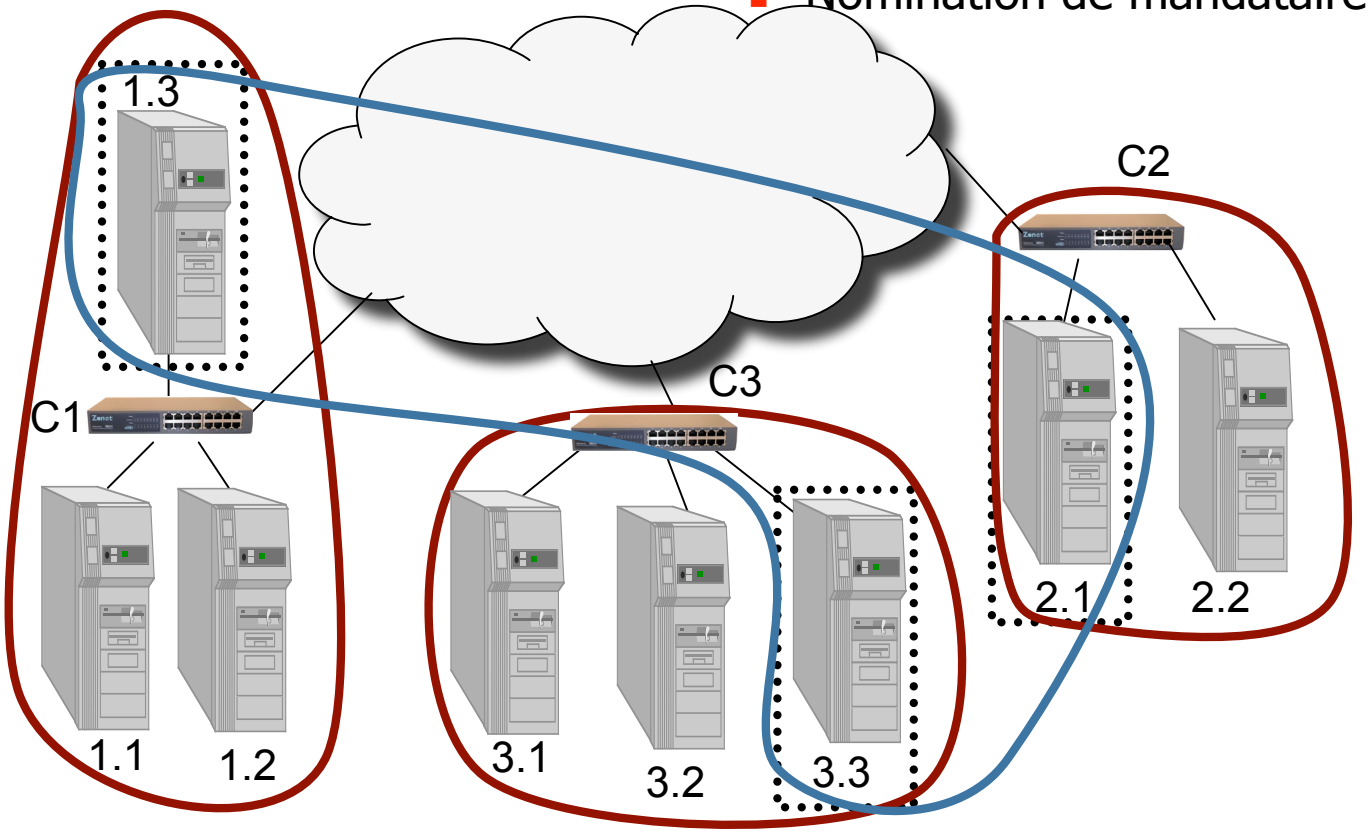
- Plateforme de programmation d'application P2P
  - Ensemble de protocoles
- Un pair
  - Identifiant unique (ID)
  - Adresse indépendante de la localisation physique
  - Plusieurs protocoles d'accès réseau (TCP, HTTP, etc)

http://www.irisa.fr/paris/Juxmem

Tolérance aux fautes : organisation hiérarchique

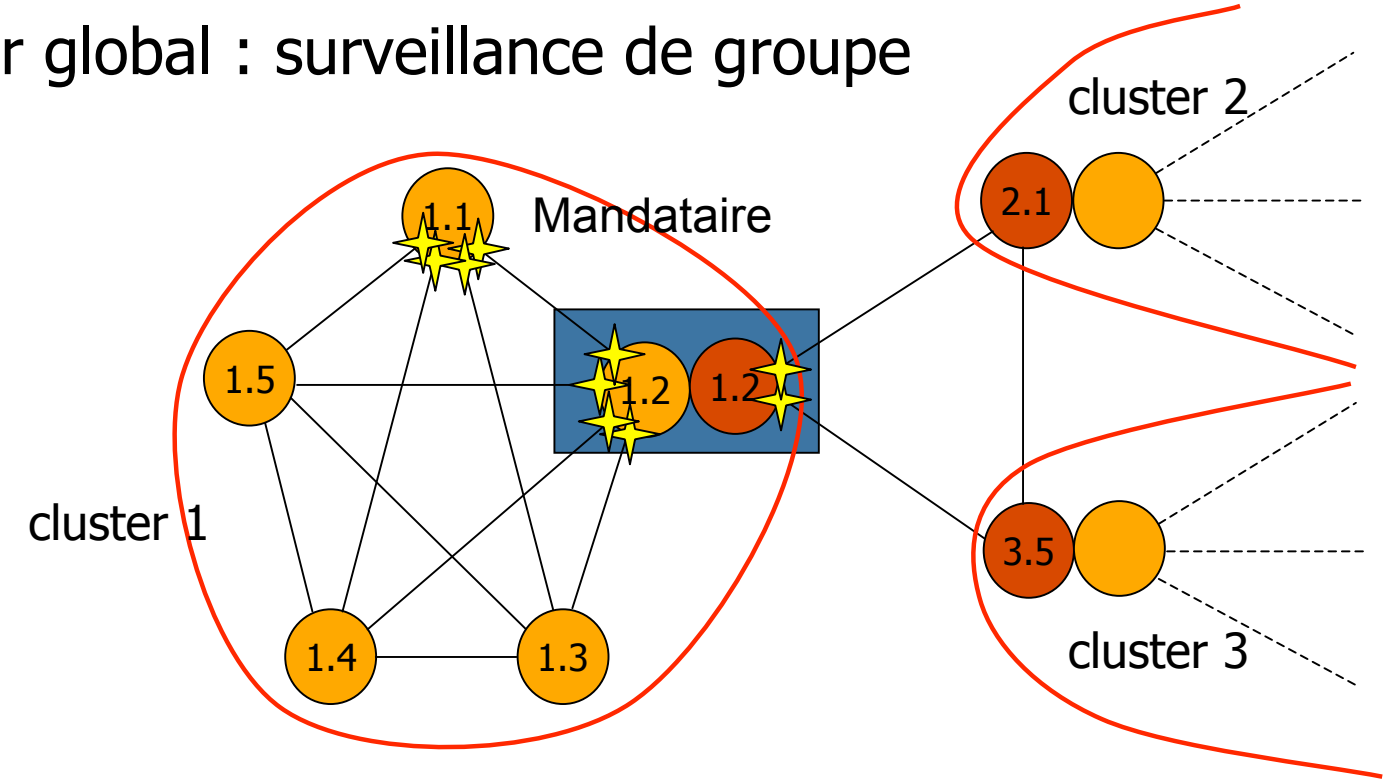
- 1 groupe local / cluster
- 1 groupe global

- Composition du groupe global
  - Représentant dans chaque groupe local = mandataire
  - Nomination de mandataire



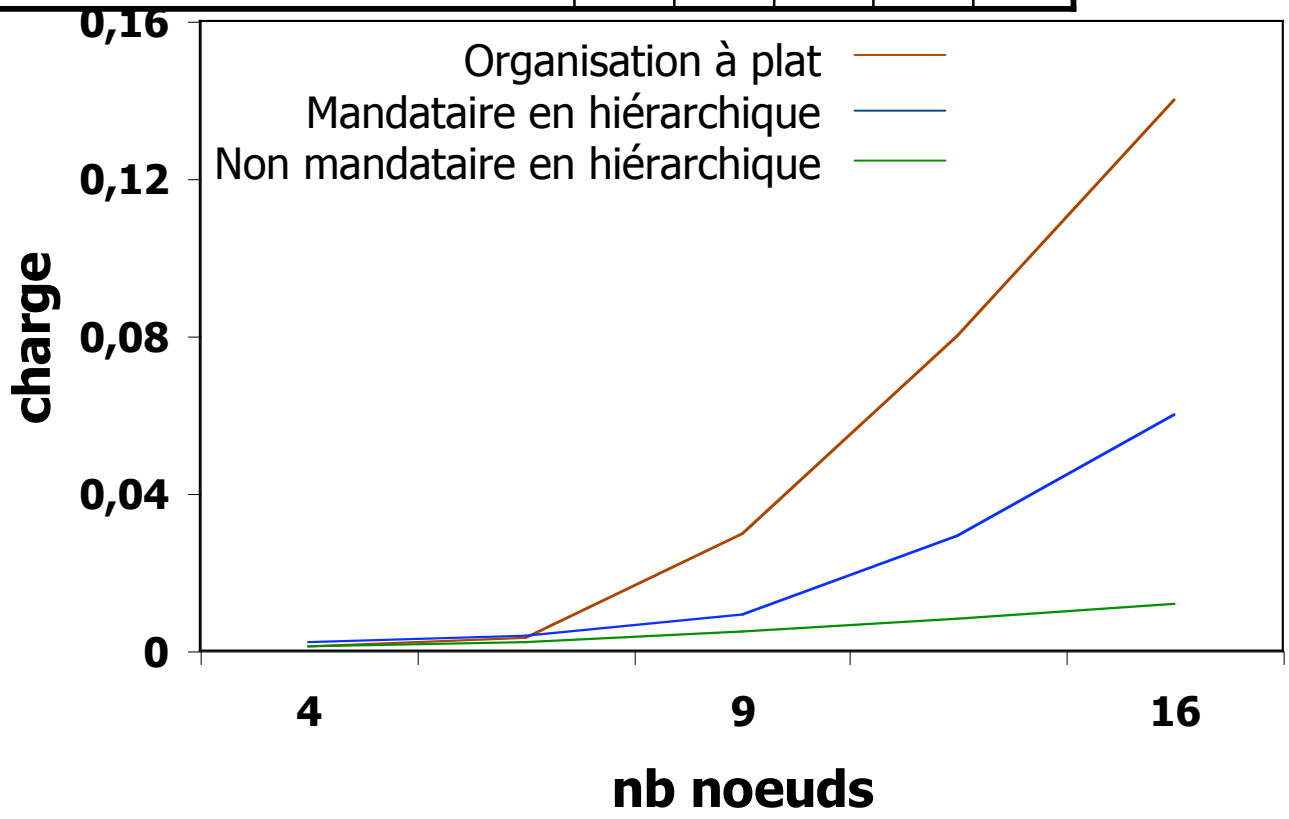
Rôle du mandataire

- Permet la connexion du groupe local avec le reste du système
- Détecteur local : surveillance de nœud
- Détecteur global : surveillance de groupe



Performances

Nb de nœuds	4	6	9	12	16
Nb de nœuds par groupe local	2	3	3	4	4
Nb de groupes locaux	2	2	3	3	4

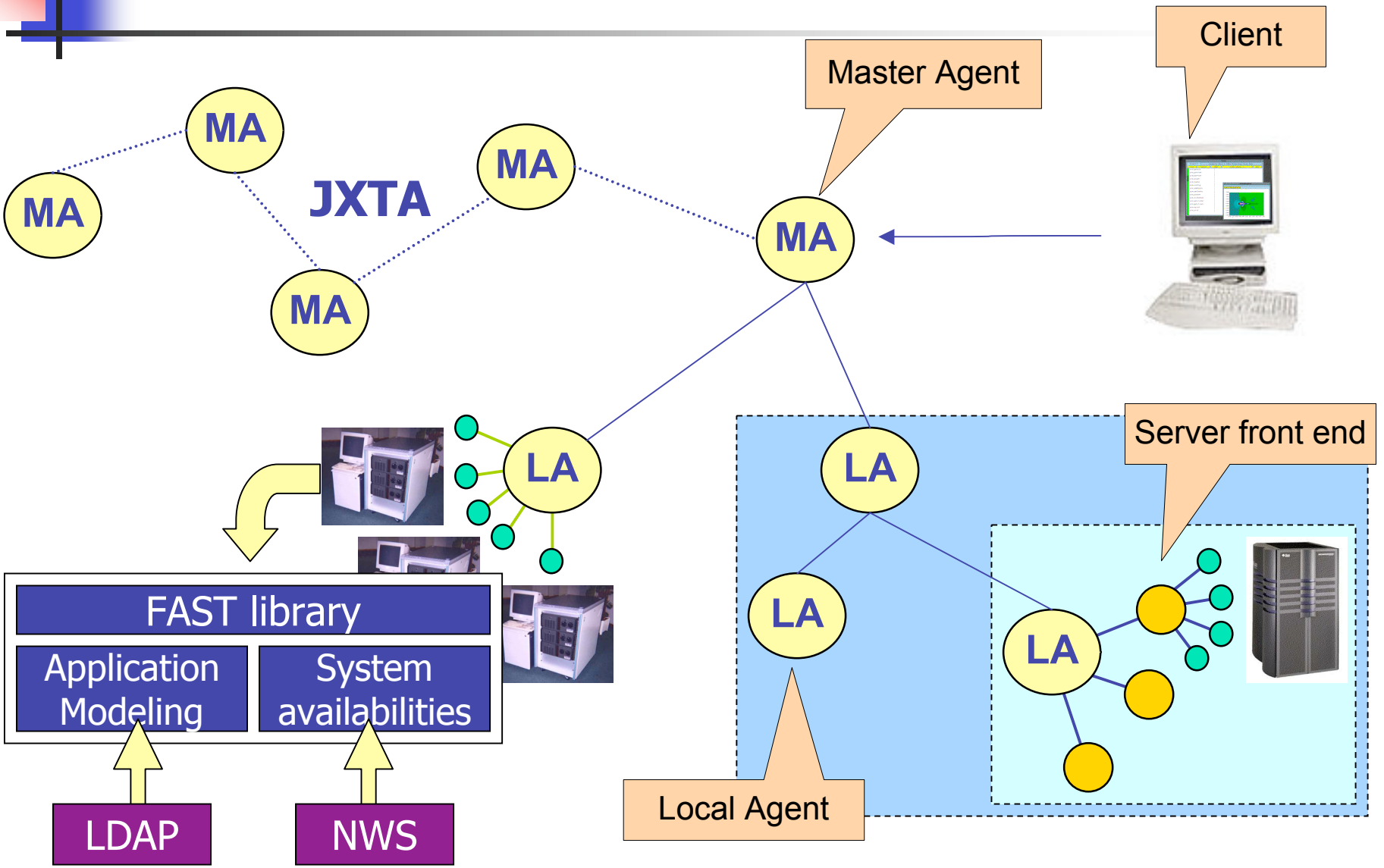


DIET

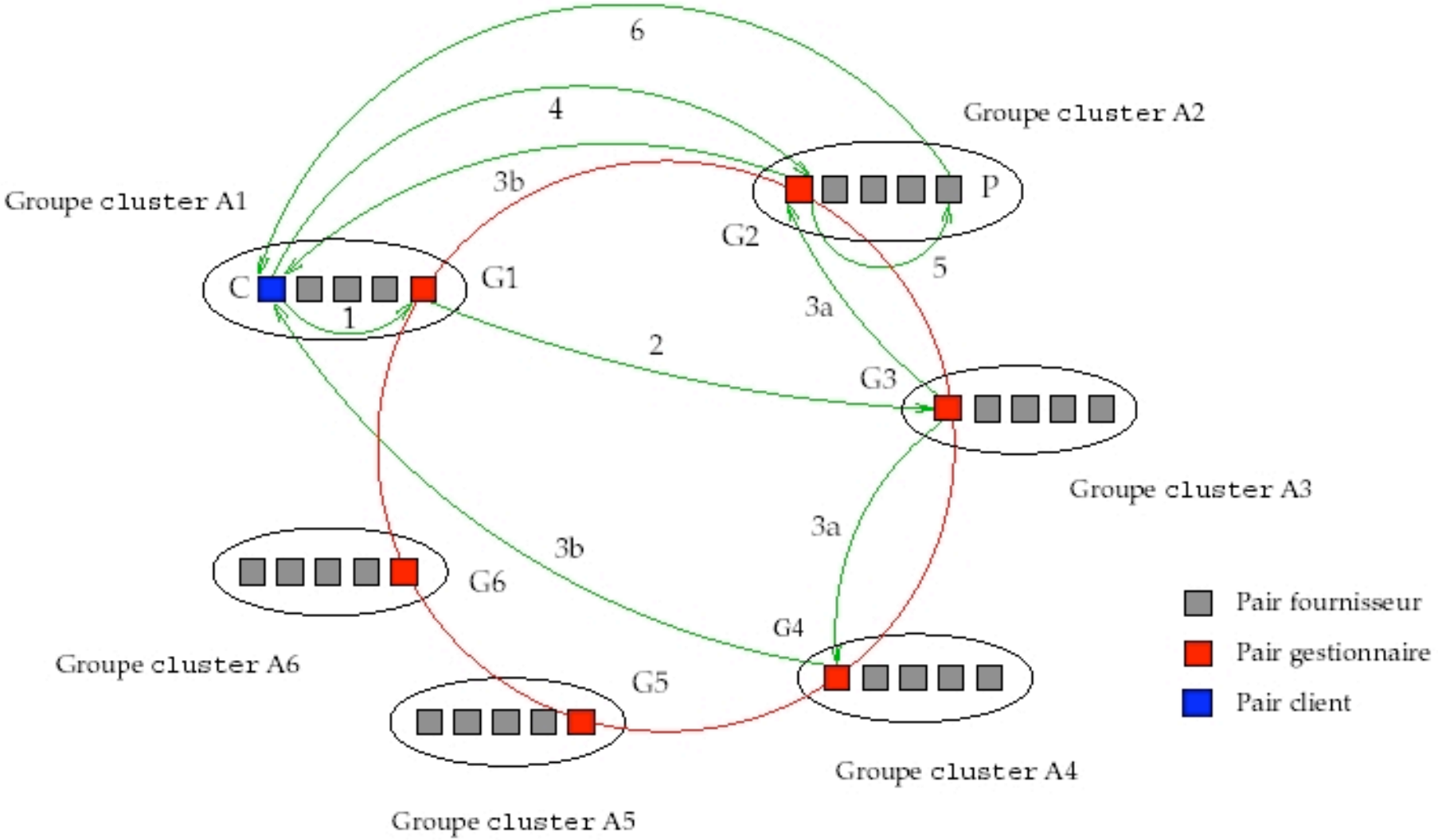
DIET

- Our goals
  - To develop a toolbox for the deployment of environments using the Application Service Provider (ASP) paradigm with different applications
  - Use as much as possible public domain and standard software
  - To obtain a high performance and scalable environment
  - Implement and validate our more theoretical results
    - Scheduling for heterogeneous platforms, data (re)distribution and replication, performance evaluation, algorithmic for heterogeneous and distributed platforms, …
- Based on CORBA, NWS, LDAP, and our own software developments
  - FAST for performance evaluation,
  - LogMgr for monitoring,
  - VizDIET for the visualization,
  - GoDIET for the deployment
- Several applications in different fields (simulation, bioinformatic, …)
- Release 2.0 available on the web
- ACI Grid ASP, RNTL GASP

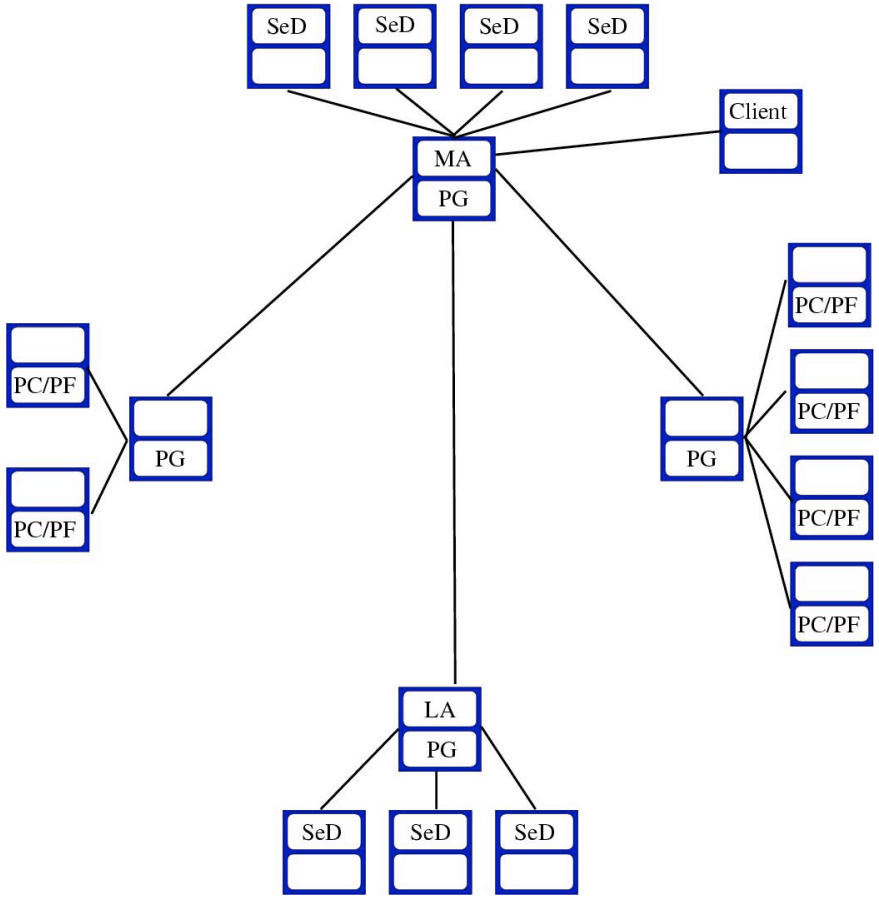
Architecture DIET

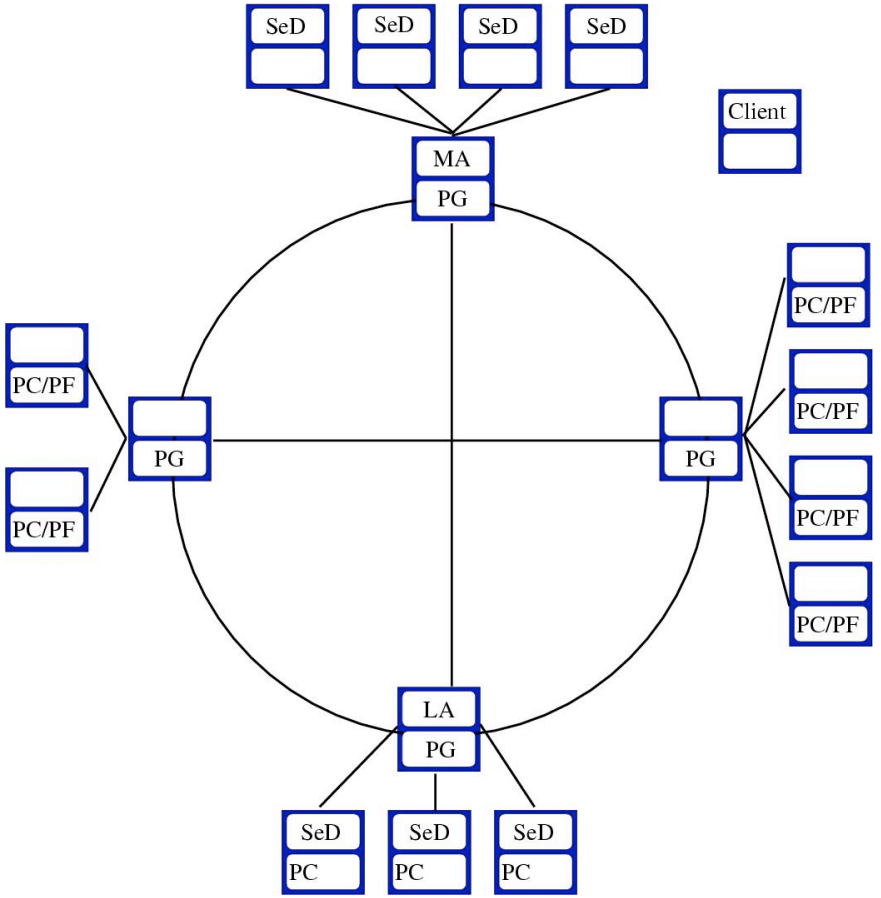


Architecture JuxMem



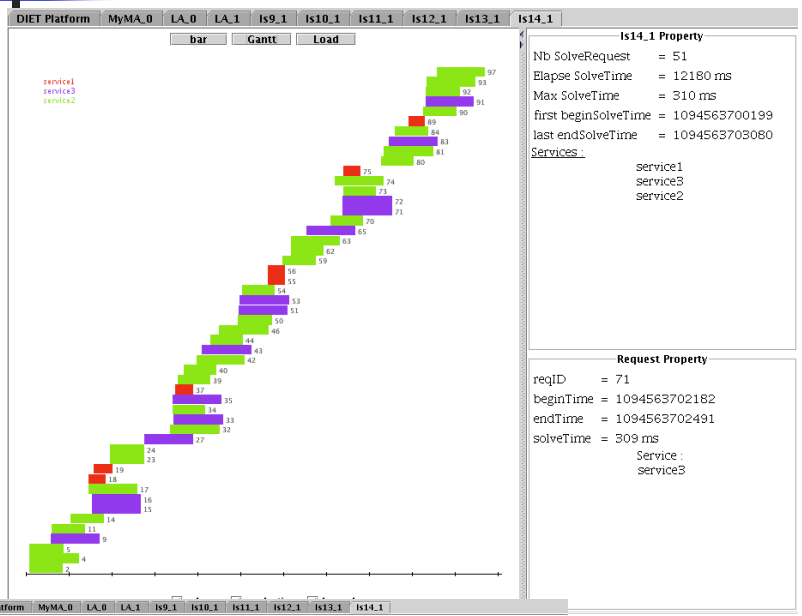
Cohabitation DIET / JuxMem

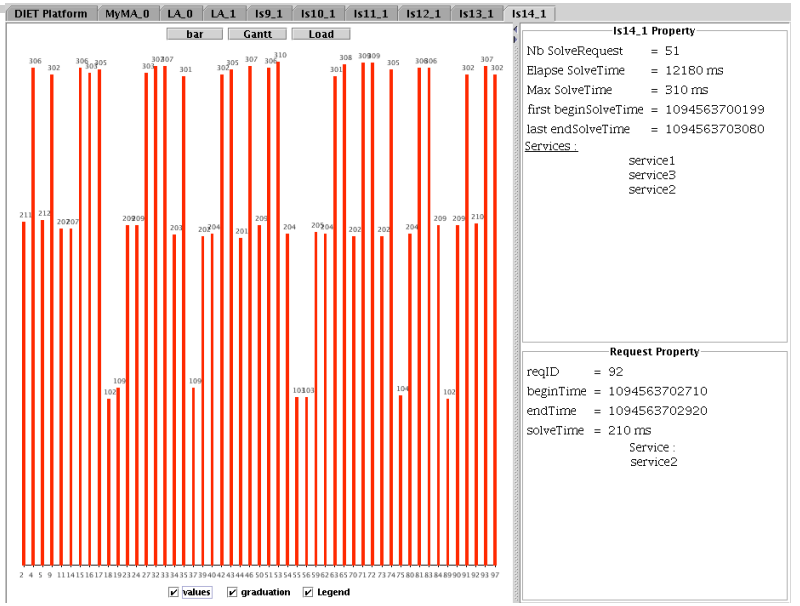


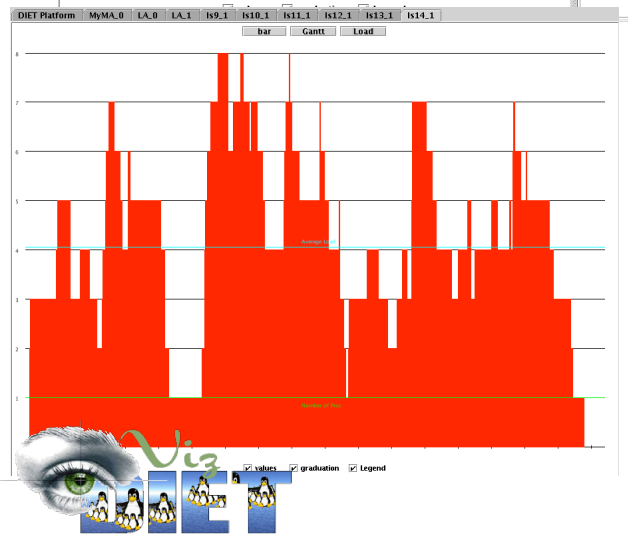


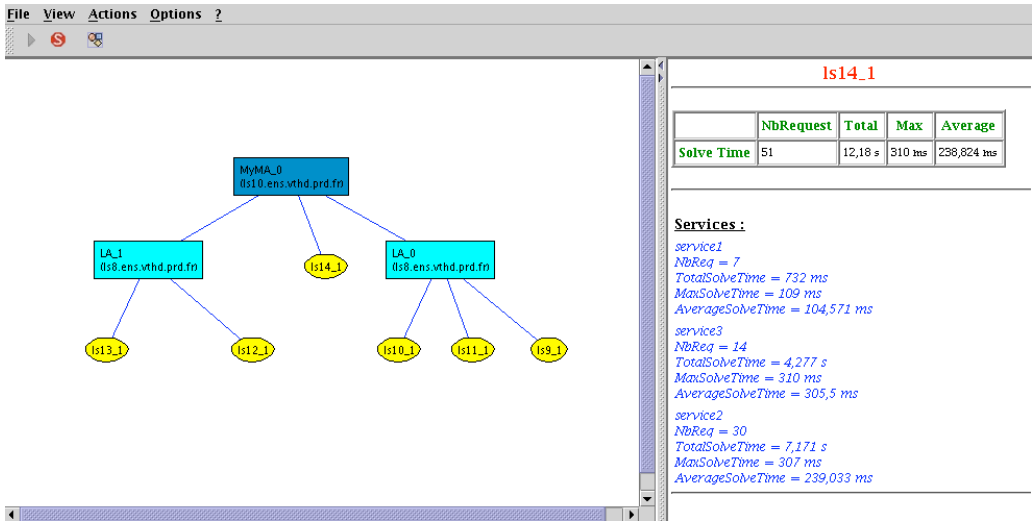
- Données répliquées sur les grappes de la grille
- Implémentation multi-protocoles (réplication, cohérence)

Screenshot


DIET Platform
MyMA_0
LA_0
LA_1
ls9_1
ls10_1
ls11_1
ls12_1
ls13_1
ls14_1
bar
Gantt
Load
ls14_1 Property
Nb SolveRequest = 51
Elapse SolveTime = 12180 ms
Max SolveTime = 310 ms
first beginSolveTime = 1094563700199
last endSolveTime = 1094563703080
Services :
service1
service3
service2
Request Property
reqID = 71
beginTime = 1094563702182
endTime = 1094563702491
solveTime = 309 ms
Service :
service3

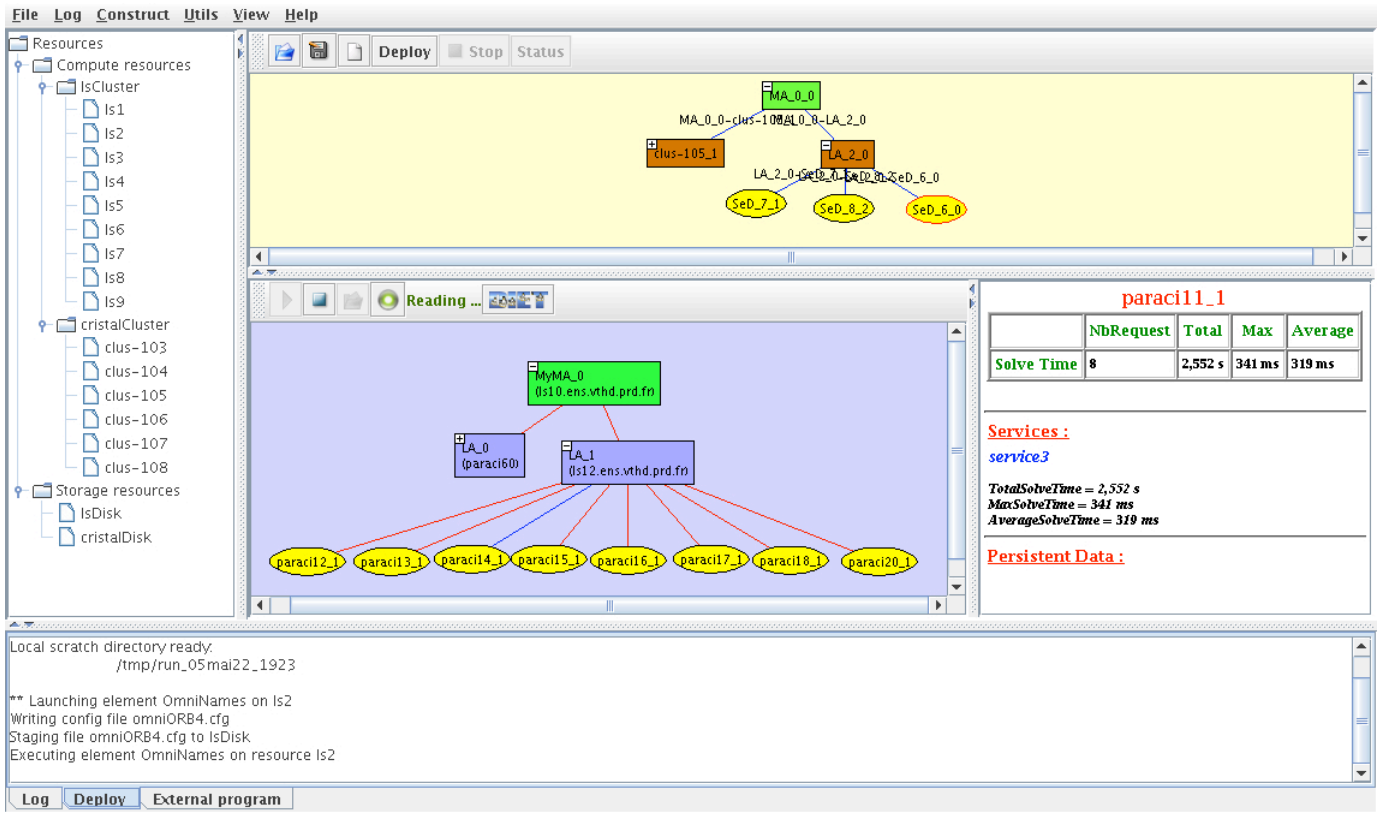

DIET Platform
MyMA_0
LA_0
LA_1
ls9_1
ls10_1
ls11_1
ls12_1
ls13_1
ls14_1
bar
Gantt
Load
ls14_1 Property
Nb SolveRequest = 51
Elapse SolveTime = 12180 ms
Max SolveTime = 310 ms
first beginSolveTime = 1094563700199
last endSolveTime = 1094563703080
Services :
service1
service3
service2
Request Property
reqID = 92
beginTime = 1094563702710
endTime = 1094563702920
solveTime = 210 ms
Service :
service2
values
graduation
Legend

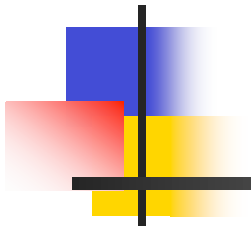

DIET Platform
MyMA_0
LA_0
LA_1
ls9_1
ls10_1
ls11_1
ls12_1
ls13_1
ls14_1
bar
Gantt
Load
values
graduation
Legend


File
View
Actions
Options
?
ls14_1
NbRequest
Total
Max
Average
Solve Time
51
12,18 s
310 ms
238,824 ms
MyMA_0
ls10.ens.vthd.prd.fr
LA_1
ls8.ens.vthd.prd.fr
ls14_1
LA_0
ls8.ens.vthd.prd.fr
ls13_1
ls12_1
ls10_1
ls11_1
ls9_1
Services :
service1
NbReq = 7
TotalSolveTime = 732 ms
MaxSolveTime = 109 ms
AverageSolveTime = 104,571 ms
service3
NbReq = 14
TotalSolveTime = 4,277 s
MaxSolveTime = 310 ms
AverageSolveTime = 305,5 ms
service2
NbReq = 30
TotalSolveTime = 7,171 s
MaxSolveTime = 307 ms
AverageSolveTime = 239,033 ms

DIET Dashboard

- Current works : We designed a new tool to merge GoDIET and VizDIET
- Resource visualization before GoDIET deployment
- Interactive GoDIET
- Only one tool to manage all DIET platform.
- DIET dashboard is designed from BuilDIET project





Les équipes – avancement

Projet GRAAL

INRIA Rhône-Alpes - CNRS – ENS Lyon

- Domaine de recherche :
  - Algorithmique et ordonnancement pour plates-formes hétérogènes distribuées
- Compétences de GRAAL (relatives à GDS) :
  - Environnements de calcul sur grille de type Network Enabled Servers (plate-forme DIET)
  - Ordonnancement des calculs
  - Algorithmes et outils de déploiement hiérarchique
- Personnels impliqués directement dans GDS :

Permanents	**Doctorants**	**Ingénieurs**
▪ Eddy Caron (MDC ENS Lyon) ▪ Frédéric Desprez (DR INRIA)	▪ Pushpinder-Kaur Chouhan (MESR) ▪ Cédric Tedeschi	▪ Raphael Bolze (INRIA) ▪ Holly Dail (INRIA)

Projet PARIS

INRIA Rennes - CNRS - Université de Rennes 1
INSA de Rennes - ENS Cachan / Bretagne

- Domaine de recherche :
  - Modèles de programmation et supports exécutifs pour les systèmes parallèles et distribués (grappes, grilles, P2P)
  - Applications : simulations numériques distribuées
  - Gestion de données pour grilles de calcul et systèmes P2P

- Compétences de PARIS (relatives à GDS) :
  - Partage de données à grande échelle (plate-forme JuxMem, basée sur la plate-forme JXTA de Sun Microsystems)
  - Protocoles de cohérence des données

- Personnels impliqués directement dans GDS :

Permanents	**Doctorants**	**Stagiaires**
▪ Gabriel Antoniu (CR INRIA) ▪ Luc Bougé (Prof. ENS Cachan, Antenne de Bretagne)	▪ Mathieu Jan (INRIA/Région Bretagne) ▪ Sébastien Monnet (ACI MD - GDS) ▪ Loic Cudennec (Sun Microsystems/ Région Bretagne)	▪ Jean-François Deverge (Master 2, IFSIC) ▪ David Noblet (UNH, USA) ▪ Chester Tse (MIT, USA) ▪ Arvind Saraf (MIT, USA)

Projet REGAL

INRIA Rocquencourt - LIP6

- Domaine de recherche :
  - Adaptation des systèmes aux environnements à large échelle
  - Architectures cibles : GRID / P2P
- Compétences de REGAL (relatives à GDS) :
  - Algorithmique répartie
  - Tolérance aux fautes (observation et détection)
  - Réplication du code et des données
- Personnels impliqués directement dans GDS :

Permanents	**Doctorants**	**Stagiaire**
L. Arantes (MDC Paris 6) P. Sens (Prof. Paris 6)	Marin Bertier (MESR)* Jean-Michel Busca (INRIA) Fabio Picconi (ACI MD - GDS) Julien Sopena (MESR)	Daniel Myers (Fullbright, USA)

* Actuellement MdC INSA de Rennes, projet PARIS

Etat du projet à T0+24

- 10 réunions GDS entre septembre 2003 et novembre 2005
- Visites des doctorants (sur 2-3 jours)
  - 3 visites de Mathieu Jan (PARIS) chez GRAAL
  - 2 visite de Fabio Picconi (REGAL) chez PARIS
  - 2 visites de Sébastien Monnet (PARIS) chez REGAL
- Logiciels en cours de développement :
  - PARIS : plate-forme JuxMem, outil de déploiement pair-à-pair JDF
  - REGAL : simulateur de systèmes à large échelle LS3, système de fichiers P2P Pastis
  - GRAAL : plate-forme DIET, outil de déploiement GoDIET et de visualisation VizDIET
- Résultats obtenus
  - Détecteur hiérarchique de fautes (REGAL) intégré dans JuxMem (PARIS)
  - Expérimentations préliminaires DIET (GRAAL)/JuxMem (PARIS)
- Travaux en cours
  - Premier prototype DIET/JuxMem, validation à l'aide l'application TLSE
  - Tests d'extensibilité : protocoles de cohérence tolérants aux fautes, algorithmes d'allocation d'espace de stockage

Etat du projet à T0+24 (suite)

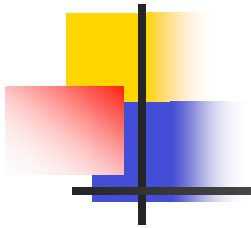
- Plate-forme matérielle
  - GRAAL : 13 nœuds + plate-forme Grid'5000 Lyon (112 noeuds)
  - PARIS : noeuds GDS intégrés dans la plate-forme Grid'5000 (264 nœuds bi-processeurs)
  - REGAL : Mise en place d'une plate-forme d'émulation de système à large échelle : 20 machines (22 nœuds bi-processeurs de calcul, 18 mono-processeurs avec 4 ralentisseurs de réseau dummynet)

- Visites extérieures
  - Visite de G. Antoniu chez *Sun Microsystems* à Santa Clara (équipe JXTA) – novembre 2003
  - Séjours de Mathieu Jan (3 mois) et de Loïc Cudennec (2 mois) chez *Sun Microsystems* - 2005
  - Visite de F. Picconi à *l'Université de Rice* (équipe de P. Drushel) – juillet 2004
  - Séjour de F. Picconi l'Univeristé de Rutgers (équipe de L. Iftode) – Janvier-Février 2005

- Invités extérieurs
  - Stage de 9 semaines de David Noblet (*Université du New Hampshire*) chez PARIS – été 2004
  - Visite de P. Hatcher (*Université du New Hampshire*) chez PARIS – mai 2004
  - Visite d'André Schiper (*EPFL*) chez REGAL – juin 2004

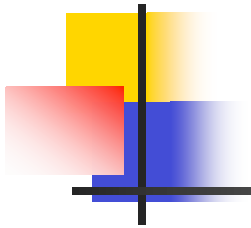
Publications GDS à T0+24

- Un chapître de livre à paraître (*Future Generation Grids*, Springer)
- Journaux internationaux (2)
  - 2005 : Kluwer Journal of Supercomputing, Scalable Computing: Practice and Experience, Concurrency and Computation : Practice and Experience
  - 2005 : Journal of Parallel and Distributed Computing
  - 2005 : International Journal of High Performance Computing Applications
  - Soumission : Journal of Grid Computing
- Conférences internationales (8)
  - 2003 : PACT (WIP), IPDPS
  - 2004 : ICPADS, CCGRID, SBAC, Euro-Par
  - 2005 : HPCC, Europar (2), e-Science Grid
- Colloques internationaux (6)
  - 2003 : AGRIDM (PACT)
  - 2004 : AGRIDM (PACT), HCW, GP2PC (CC-GRID), WGAPI (GGF-12), WGC
- Communications nationales (5)
  - 2003 : RenPar
  - 2004 : Ecole DRUIDE (3), GridUSe
  - 2005 : CFSE, Journées CDUR



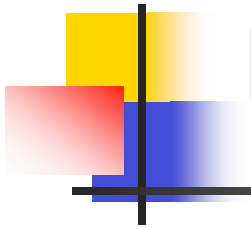
Collaborations nationales

- ACI GRID DataGraal (animation)
  - Ecolé thématique DRUIDE 2004 : Distribution de données à grande échelle (CNRS, INRIA, GDR ARP, Univ. Rennes 1), Le Croisic, mai 2004
- ACI MD
  - GdX – expérimentations GDS/GdX prévues pour 2005/2006 (émulation)
  - MDP2P - workshop sur la gestion de données P2P (mars 2005)
  - Gédéon (protocole de réplication)
- AS : Algorithmique Distribuée et Applications
  - Journées thématiques « algorithmique distribuée et applications », Porquerolles, septembre 2004
- Grid'5000
  - Déploiement et tests à grande échelle multi-sites



Collaborations internationales

- Partenaires académiques
  - AIST, Tokyo (partage transparent P2P à grande échelle)
  - Réseau Européen d'Excellence CoreGRID / Université de Pise (stockage de données sur grille)
  - Univ. Rice (système P2P)
  - Univ. Libre d'Amsterdam (réplication adaptable)
  - EPFL (détection de fautes)
- Partenaires industrielles
  - Sun Microsystems : JuxMem/JXTA/ Grid'5000
    - Financement d'une thèse à l'IRISA démarrée en 2005 (Loïc Cudennec)



Pour plus de détails…

- Site web : http://www.irisa.fr/GDS/
  - Programme des réunions
  - Présentations en ligne
  - Visites
  - Publications
  - Archive liste de diffusion : acimd-gds@irisa.fr